Aligning Items and Achievement Levels: A Study Comparing Expert Judgments

Pamela Kaliski, The College Board

Kristen Huff, Regents Research Fund

Carol Barry, The College Board

CollegeBoard
inspiring minds

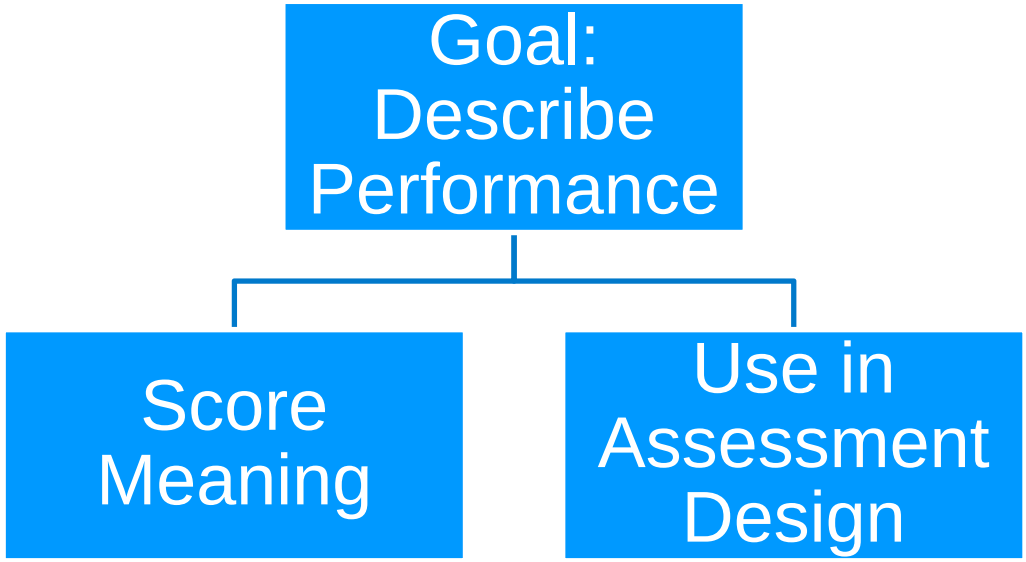
Overview of presentation

- Goal of developing achievement level descriptions
- Importance of aligning items to achievement levels
  - Identifying difficulty drivers to improve item-ALD alignment
- Current study: Piloted a methodology to work with SMEs to identify difficulty drivers
  - Results and discussion of next steps

CollegeBoard
inspiring minds

Achievement Level Descriptions in K-12 Testing

- ALDs = description of what students know and can do at a particular level
  - Describe how performance differs for students in each category



e.g., Huff & Plake, 2010

ALDs and item development

- When aim is to reliably classify students into performance levels, must design items capable of discriminating student performance
  - Write items to intended ALD levels (e.g., AP ALDs for 1, 2, 3, 4, 5)
- What item design features relate to ALDs?
- If items are designed w/o goal of classification and alignment, there is risk that the student performance on an exam will not align with intended interpretation
  - Especially important for tests with MC items (e.g., AP exams)

e.g., Schneider et al., 2010

CollegeBoard
inspiring minds

Difficulty drivers

- Alignment between items and ALDs is difficult to achieve! We need to understand what item features are driving the difficulty/complexity of an item
- Identifying **difficulty drivers** for use in item development is the key to achieving this alignment.
- Can you really predict item difficulty?
  - NO. The relationship between item difficulty and achievement level will never be perfect, but we certainly can identify item features that relate to item difficulty
  - Should be a negative relationship between p-values and performance levels

e.g., Barry & Huff, 2009

CollegeBoard
inspiring minds

Relationship between item difficulty and achievement level

Purpose of study

To pilot a new methodology for identifying difficulty drivers with SMEs for AP US History. This exam is currently undergoing a redesign project using evidence-centered assessment design.

RESEARCH QUESTIONS:

1) What item features make items more or less difficult for this AP US History exam?

2) Did the methodology employed in this particular item alignment workshop elicit relevant difficulty drivers?

CollegeBoard
inspiring minds

Methods: Participants and Pre-meeting assignment

- 4 strongly recommended SMEs in United States history were recruited to participate in a 1.5 day workshop.
  - 2 high school, 2 college
- Pre-meeting assignment, SMEs were asked to classify AP US History items to ALD level
  - 18 items
  - Drafted ALDs
  - Spreadsheet for homework responses

CollegeBoard
inspiring minds

Methods: 1.5 day workshop

- Detailed review of 9 of the 18 items
  - The intended ALD level and skills were revealed.
  - Results of the SME classifications were presented, and the SMEs discussed their rationales with one another.
  - Presentation of actual student performance data from the item pilots.
    - Item difficulty data (percentage of students who got the item correct)
    - Opportunity to learn data (percentage of teachers who currently teach the content and/or skills necessary to correctly answer the item) were presented.

CollegeBoard
inspiring minds

Methods: 1.5 day workshop, continued

- After sufficient discussion, the research question "What features make items more or less difficult?" was projected on a screen in front of the room
- A list of "difficulty drivers" was documented and saved.
- Evaluation survey
- Follow up assignment after meeting to confirm what was documented during the meeting

Results: Draft taxonomy of difficulty drivers

- Item characteristics
  - Degree of scaffolding
  - Word count
  - Vocabulary unrelated to target of measurement
  - Number of components to an item
- Target of measurement characteristics
  - Degree of familiarity
  - Plausibility of distracters

Discussion

- "Difficulty drivers" that emerged using this method with SMEs were not really related to historical sources of cognitive complexity
- Underscores how difficult this process can be, particularly for this domain!
  - *How can we better elicit difficulty drivers related to the domain?*
- Evaluation survey was helpful—SMEs valued experience and this research, were clear on goals of the meeting
  - Indicated that this is a difficult task

CollegeBoard
inspiring minds

Possible suggestions for future

- Add a learning scientist
- Add an additional piece of pre-meeting work, to have SMEs beginning to think about difficulty drivers in advance
- Increase the length of the meeting
- ALD workshop
- WHAT ELSE?

CollegeBoard
inspiring minds